Exploring the Relationship between Perceived Health and Actual Health Across Income Categories

Filippo Ravalli and Nicholas Rubertone

Applied Data Mining

Project 3

April 19, 2021

Github Link: https://github.com/filipporavalli/Project-3-Data-Mining

Abstract

The motivation for this report stems from existing literature which shows that lower income contributes to poorer health outcomes and that perceived health can be a potent predictor for short-term mortality. We therefore decided to examine the relationship between perceived health and actual health across different levels of income. Using data pulled from the CDC-led National Health Interview Survey (NHIS), we imputed weights from the Global Burden of Disease study to medical conditions which overlapped in both datasets to create a composite health score for each individual. To identify patterns, a linear regression model was run as a form of data mining, which resulted in findings of discrepancies between actual health and perceived health being dependent on one's income category. Further exploration found that increasing income generally resulted in an increasing perception of being in better health and higher accuracy in perceived health for those who actually had good health. For lower income individuals, the converse was true in that those with poorer health were more accurately labeling themselves as having poor health. Differences between perceived and actual health were also smaller as income decreased. To improve the findings in this report, we would have liked to have a dataset with a greater breadth of medical conditions that would allow for a more accurate composite health score feature. Although the engineered feature proved useful in our analysis, limitations in the imputed weights and limited overlap between the two datasets meant that the composite health score may not have been a comprehensive reflection of an individual's health. Despite this, our results were corroborated with existing studies that found associations between income and perceived health.

Introduction

Low income has long been associated with poorer health outcomes and increased mortality due to limited access to quality healthcare and poor social living conditions[1]. The relationship between perceived health and income, however, has received less attention. Research has shown that perceived health can be a strong predictor for short-term mortality[2], so understanding if there is a relationship between income and perceived health can help influence public health strategies and policies aimed at improving both one's actual health and general perception of health. For example, any disparities in perceived health due to income could lead to certain groups not seeking out medical care due to a perception of being in good health, when in fact their health is poor. Public health work could subsequently focus on ensuring individuals stay up to date on medical visits and follow a healthy lifestyle. The aim of this project is to explore the relationship between perceived health and actual health across different levels of income. Based on our findings, which showed an income-dependent pattern in perceived health choice, we also examined if the accuracy of perceived health matching actual health differed across income groups.

In tackling this problem, we initially used data from the National Health and Nutrition Examination Survey (NHANES), a personal interview based survey conducted by the Centers for Disease Control and Prevention (CDC) that aims to capture the general health condition of Americans. However, due to the limited number of participants (<9,000) and a high degree of missingness in questions pertaining to medical conditions, we turned to the National Health

[1] Adler, Nancy E., and Katherine Newman. "Socioeconomic Disparities in Health: Pathways and Policies." *Health Affairs (Project Hope)* 21, no. 2 (April 2002): 60–76. https://doi.org/10.1377/hlthaff.21.2.60.

[2] Kim, Jae-Hyun, and Jang-Mook Kim. "Subjective Life Expectancy Is a Risk Factor for Perceived Health Status and Mortality." *Health and Quality of Life Outcomes* 15, no. 1 (October 2, 2017): 190. https://doi.org/10.1186/s12955-017-0763-0.

Interview Survey (NHIS), another CDC-led personal interview survey, which had a larger sample size and only a small amount of missing data. The primary question of interest remained unchanged, as we focused on examining the association between income and perceived health with respect to actual health.

Dataset Background

The primary dataset was taken from NHIS as it provided both the medical and demographic data necessary to investigate the question. This survey is used by the CDC to collect data on a broad range of topics including health status, health care access, and progress toward achieving national health objectives through personal interviews[3]. The survey is conducted annually and aims to conduct around 30,000 sample adults and 9,000 sample child child interviews. Examples of questions asked during the survey include demographic information (e.g., income, ethnicity, age, etc.) and health-related questions (e.g., "Ever been diagnosed with bone cancer," "Personal health status," or "Ever had surgery"). For this report, we pulled sample adult data from 2019 directly from the CDC website. The total number of subjects interviewed was 31,997 and the total number of features was 534.

Although NHIS data provides many variables on personal health, it lacks an objective composite health variable that defines the general health of an individual. To examine actual health, we therefore used a secondary dataset which provided weights associated with different conditions and diseases from the 2019 Global Burden of Disease Study. Briefly, the Global Burden of Disease Study is an epidemiological study conducted by the Institute for Health Metrics and Evaluation (IHME) that shows where certain diseases are most prevalent across the world and estimates the total burden and disability associated with particular diseases[4]. Since

[3] "NHIS - National Health Interview Survey," April 14, 2021. https://www.cdc.gov/nchs/nhis/index.htm.
[4] "About the Global Burden of Disease." Accessed April 18, 2021. https://www.thelancet.com/gbd/about.

every disease brings a different level of disability, each condition and its severity are assigned a particular weight between 0 and 1, 0 representing no burden and 1 representing death. We therefore used the weights to create an estimate of actual health for individuals using conditions that overlapped in both datasets. Data was pulled directly from the IHME for the 2019 study. The data provided weights for 2,117 different diseases or associated severities (severe stages of a disease have a different burden than early stages/asymptomatic stages).

Data Processing

The NHIS dataset was subsetted to contain only 35 features of interest, including income category (possible choices shown in figure 1), perceived health status, and 33 specific features for different medical conditions (Appendix Table 1). The majority of these conditions were chronic, including 25 different types of cancer. These medical conditions were selected based on the secondary dataset, as described below. Each of the medical condition features indicate if an individual has ever been diagnosed with that specific feature (Possible answers were: "Yes," "No," "Don't Know," "Refused," or "Not Ascertained"). All of the features associated with cancer had a high degree of missingness, though further investigation of the data codebook highlighted that a missing value for these features corresponded with a response of "No." Only one feature, "History of abdominal pain in the past 90 days," had truly missing data with 39.78% of the feature's values being labeled as missing. The primary feature of interest, perceived health status, was an ordinal variable (1="Excellent," 2="Very good," 3="Good," 4="Fair," 5="Poor", 7= "Refused", 9= "Don't Know"). Rows with responses marked as "Refused" or "Don't Know" were removed (n=15). Both income data and perceived health status had no missing data.

The primary motivation for the secondary dataset was to engineer a composite health feature. Using the IHME data, we selected conditions in the NHIS data that had corresponding weights available. Based on participants' answers to the questions for each medical condition, we manually imputed the weight values for individuals who answered “Yes” to ever being diagnosed for each specific condition. For diseases that had different weights for different degrees of severity in the IHME data, the average was taken to produce a single weight for each disease that was subsequently combined with the NHIS dataset. This method produces obvious limitations, as it may downplay the increased weight of severe stages of diseases or exaggerate the low severity of a disease’s early stages; however, the NHIS data lacked any mention of disease severity for the conditions it contained. For individuals who answered “No,” or had a missing value to a specific question, a weight of 0 was imputed for that condition. Finally, the sum across the imputed weights for all medical conditions was calculated, resulting in each individual having a composite health score that we used as a proxy for objective health. The range of score values was 0 – 2.411, with score of 0 indicating perfect health according to the diseases mentioned in the survey.

In addition to averaging across different severities for an individual disease, the weights themselves may raise data quality concerns. Defining how much disability a particular condition may bring is not an exact science, so the health score we created may not be a perfect reflection of the individual’s actual health. Despite this limitation, we believe that our health score was able to provide relatively accurate estimates for the objective, general health of each individual.

To examine the relationships found in our algorithm, we also created score quantiles that corresponded with the perceived health status responses, observed in Table 1. This crude approximation placed individuals into 5 health status categories based on their actual health score and allowed comparison between individuals’ perceived health status and their objective health status.

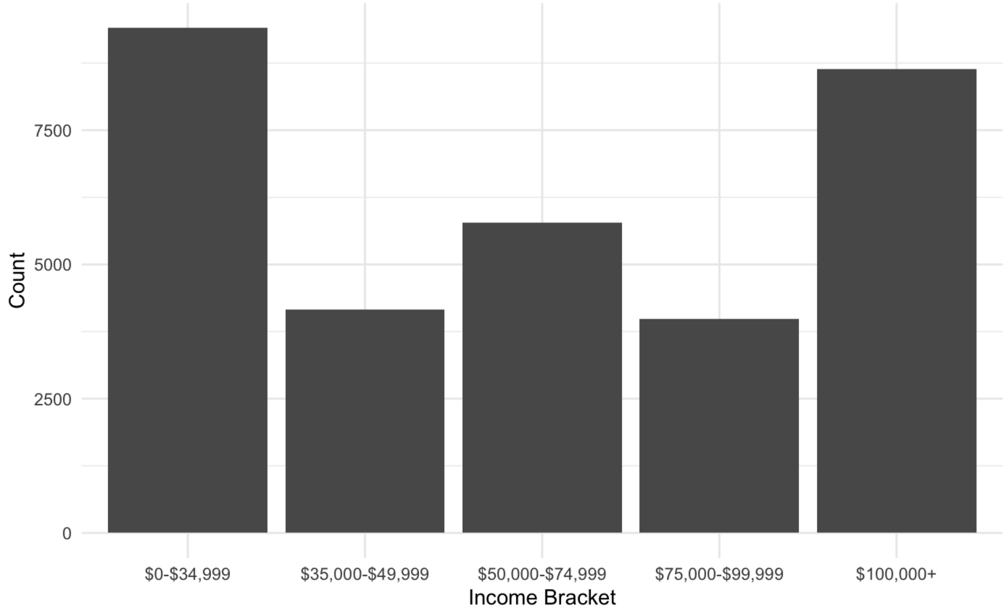
Table 1: Score Quantile & Corresponding Health Status

Quantile	Score Range	Health Status
≤ 20%	0.000-0.046	Excellent
20% to 40%	0.046-0.332	Very Good
40% to 60%	0.332-0.436	Good
60% to 80%	0.436-0.664	Fair
80% to 100%	0.664-2.411	Poor

Exploratory Data Analysis

Initial visualization of the data highlighted some patterns in the distributions of important variables. Figure 1 shows the distribution of income categories across the entire dataset and highlights that the majority of respondents fell within the highest or the lowest income category. These values are important in framing later findings about the relative presence of each income category in different health categories, as shown in Table 2.

Figure 1: Distribution of Income Categories



As observed in Figure 2, when examining the distribution of perceived health responses, the majority of respondents indicated that they had “Very Good” or “Good” health, while very few indicated that they had “Poor” health. The number of people who indicated “Excellent” health was also very high. These results suggest that, generally, people tend to view themselves as being in good health as opposed to “Fair” or “Poor” health. The histogram of the engineered score variable (Figure 3) shows that the majority of respondents had an actual health score of 0, indicating excellent health. There were also additional peaks below a score 0.5, and the overall histogram was right skewed, with potential outliers being scores above 1.5. Due to a small number of outliers (1.0%), these records were retained.

Figure 2: Distribution of Perceived Health Responses

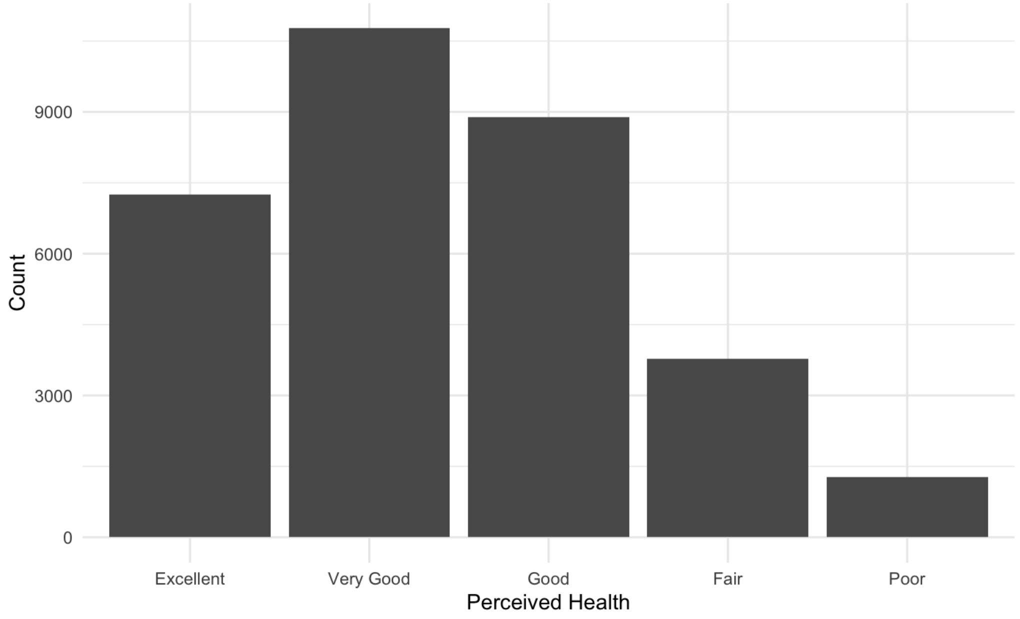
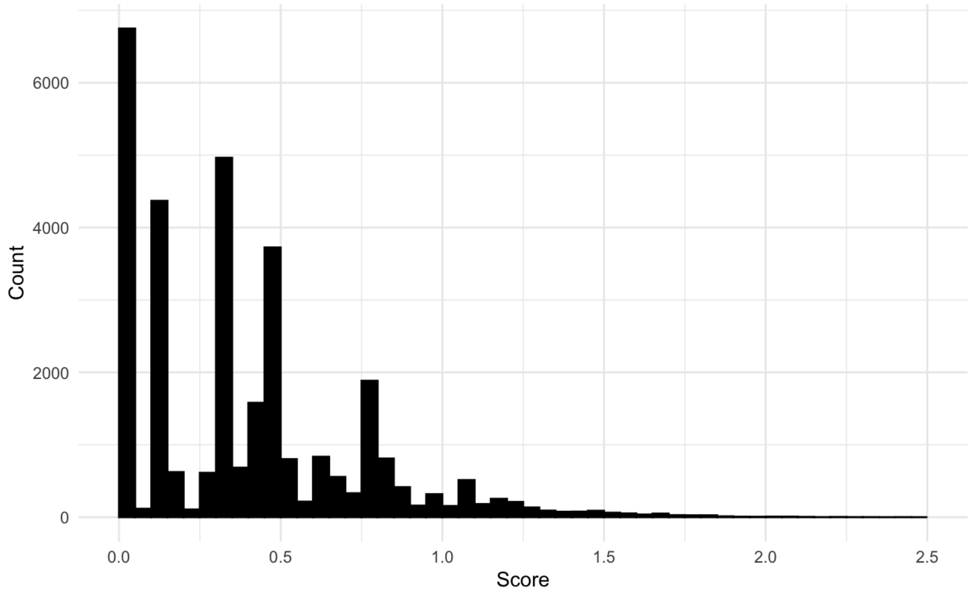


Figure 3: Distribution of Health Score



When examining perceived health across income groups (Table 2), we found that those belonging to the highest income category were most likely to perceive their health as being "Excellent," while those in the lowest income category were least likely. This trend was mirrored in the "Very Good" category as well; however, for responses of "Good" or lower, the opposite was true, as people in the highest income category were least likely to select these options.

Table 2: Perceived Health by Income

Income Category ($)	Excellent (%)	Very Good (%)	Good (%)	Fair (%)	Poor (%)
≤ 34,999	13.97	23.78	32.59	21.09	8.56
35,000 – 49,999	18.81	32.40	31.80	13.14	3.84
50,000 – 74,999	22.78	35.44	29.51	9.90	2.37
75,000 – 99,999	24.45	40.17	26.48	6.97	1.93
100,000 +	33.21	40.86	20.17	4.66	1.10

Actual health by income category (Table 3) showed discrepancies with perceived health. It shows that the proportion of people in the highest income category who truly had “Excellent” health was lower than the proportion who perceived that they had “Excellent” health. Another noticeable pattern was that the proportion of people who truly had “Poor” health was much higher across all income groups than what was perceived. When compared with Table 2, differences in perceived health versus actual health appear to be associated with income. Taking “Excellent” health as an example, a ~20% difference was observed between the lowest and highest income category for perceived health, while the same comparison for actual health showed only a ~5% difference.

Table 3: Actual Health by Income

Income Category ($)	Excellent (%)	Very Good (%)	Good (%)	Fair (%)	Poor (%)
≤ 34,999	18.56	27.17	8.08	17.68	28.51
35,000 – 49,999	20.85	31.30	7.25	18.09	22.51
50,000 – 74,999	21.72	34.44	7.08	18.10	18.66
75,000 – 99,999	21.82	36.51	7.35	19.56	14.77
100,000 +	23.26	39.36	6.36	19.11	11.90

Model Description and Model Pattern Exploration

This preliminary analysis suggested that there were indeed some differences across income categories between how people perceived their health and what the true state of their health was. In building a model to further explore these relationships, we used a linear regression where the predicting variables were the interactions between each income category and each perceived health category, and the outcome variable was the engineered health score feature. As a higher health score indicates worse health, larger coefficients in this model would indicate worse health for the given income-perceived health pair. With this model, we expected that, assuming one's perceived health is actually a good predictor of their true health, coefficients for groups that perceived their health as being worse would be larger; likewise, we expected that people of a lower income would have larger coefficients for each grouping of perceived health. We held the latter expectation for two reasons: first, Table 2 suggests that people of a lower income seem to perceive and actually have worse health on average; second, research suggests that lower income individuals experience worse health outcomes on average[5]. Assuming an accurate perception, one would therefore expect lower income individuals to still have worse health than high income individuals within each perceived health category.

Upon building our model, we found some notable trends regarding the distributions of the coefficients for each income-perceived health pair. Primarily, as perceived health worsened, the range of coefficients increased. For example, for the "Excellent" health category, the smallest coefficient was 0.158 and the largest was 0.180, but for the "Poor" health category, the smallest coefficient was 0.530 and the largest was 0.732, resulting in a range of 0.022 for the former category and 0.202 for the latter. This may suggest that income plays an increasingly important role in determining true health given the perception of one's health as this perception worsens.

[5] Chokshi, Dave A. "Income, Poverty, and Health Inequality." *JAMA* 319, no. 13 (April 3, 2018): 1312–13. https://doi.org/10.1001/jama.2018.2521.

Regarding our expectation that within each perceived health category the coefficients would be larger for lower income categories, we found this to be true in some instances, but not true in others. For example, in the “Good” and “Very Good” health categories, this trend held: the largest coefficient was for the lowest income category, and the smallest was for the highest income category, with the remaining categories falling in order. However, for the “Excellent” category, the highest income category had the largest coefficient and the lowest income category had the smallest coefficient (though the order of the other three categories did not continue sequentially), thus going against our expectation. This finding suggests that people belonging to the highest income category who perceive their health as excellent have worse health on average than people in the lowest income category with the same perceived health. As this trend was quite interesting and contrary to our expectations, we explored it further in our in-depth analysis of the differences between perceived health and health score.

Another interesting pattern that guided our later exploration of the data regarded instances in which the coefficients for a given perceived health category did not all appear directly next to each other; instead, these instances indicated that the people who labelled themselves as a worse category of health actually had better health than people who labelled themselves as having better health. For example, the coefficient for people in the highest income category who labelled themselves as having “Poor” health was smaller than the coefficients for people in the two lowest income categories who labelled themselves as having “Fair” health. Such an inconsistency further suggests a misattribution between a person’s true health and their perception of their health as stratified by income. Thus, we chose to explore in more detail how people’s perception of their health compares to their true health across income categories.

To quantitatively understand the performance of this model and determine the extent to which perceived health and income are related to health score, we compared the R-squared value of our model to models in which the perceived health or income categories were not present. In the reduced model with only income, the R-squared value was 0.029; with only perceived health, R-squared was 0.135; with both variables, the R-squared value was 0.142. The difference between these values suggests a significant increase in the explained variability by our model in comparison to the reduced models, showing the existence of a relationship between perceived health, income category, and health score. Importantly, the R-squared value being significantly smaller for the reduced model with only income category suggests that our model is dominated by the perceived health variable, but the R-squared value for the model with just perceived health not being equal to the full model shows the impact of the income variable, thus providing evidence to suggest that there is some relationship between income and health score.

Further Pattern Exploration

Based on the algorithm's patterns, we examined more closely the accuracy of participants' perceived health response with respect to their composite health score. Table 4 shows the accuracy of respondents based on their income category and true health status (i.e., for people who had a true health score corresponding to "Excellent" and belonged in the lowest income category, perceived health accuracy was 29.17%). Overall, the trends suggest that as income increases, one's accuracy also increases if the individual's true health is "Excellent" or "Very Good." This corresponds with patterns observed in Table 2, which suggested that higher income individuals perceived themselves to be in better health. Additionally, as health status decreases, higher income groups see a much larger difference in accuracy between "Excellent" and "Poor" health statuses, while the lowest income category sees a much smaller difference.

This suggests that those who belong to the lowest income category are more accurate in perceiving themselves as having poorer health, whereas those who have higher income appear to perceive themselves as being healthier than they truly are.

Table 4: Accuracy of Perceived Health Choice by Income Category and Actual Health Quantile

Income Category ($)	Excellent (%)	Very Good (%)	Good (%)	Fair (%)	Poor (%)
≤ 34,999	29.17	29.13	33.68	21.12	17.69
35,000 – 49,999	30.65	35.76	35.10	12.22	9.28
50,000 – 74,999	37.13	38.44	33.74	9.46	7.61
75,000 – 99,999	33.22	40.59	29.01	7.82	5.94
100,000 +	42.62	41.43	27.09	4.54	3.89

Seeing that the two extremes, “Excellent” and “Poor,” produced the strongest effects, we further examined the distribution of which category people were marking themselves as if they truly belonged to one of these categories. For people who truly had “Excellent” health, most correctly identified themselves as having either “Excellent” or “Very Good” health, regardless of income (Table 5). However, consistent with previous tables, increasing income resulted in increasing accuracy. Additionally, those belonging to the lowest income category indicated their health to be “Excellent,” “Very Good,” or “Good” in roughly equal amounts. Very few people listed themselves as having “Poor” health, but those with lower income still answered that they had “Poor” health at a higher rate than those with higher income. When examining the subset of the population who truly had “Poor” health, we see that increasing income resulted in decreasing accuracy (Table 6). Additionally, the overall accuracy across all income groups is much lower than for those in Table 5. People belonging to the highest income category also selected that their health was “Excellent” in higher amounts than other categories. These tables further support the notion that as income increases, individuals are more likely to accurately label themselves as having “Excellent” health, but as income decreases, individuals will more accurately label themselves as having “Poor” health.

Table 5: Distribution of Perceived Health Choice for Individuals with “Excellent” Health

Income Category ($)	**Excellent (%)**	**Very Good (%)**	**Good (%)**	**Fair (%)**	**Poor (%)**
≤ 34,999	29.17	29.86	28.65	10.20	2.12
35,000 – 49,999	30.65	35.37	27.42	5.76	0.81
50,000 – 74,999	37.13	32.75	25.42	4.14	0.56
75,000 – 99,999	33.22	40.34	21.49	4.14	0.80
100,000 +	42.62	38.84	15.85	2.19	0.50

Table 6: Distribution of Perceived Health Choice for Individuals with "Poor" Health

Income Category ($)	Excellent (%)	Very Good (%)	Good (%)	Fair (%)	Poor (%)
≤ 34,999	3.17	14.40	31.49	33.25	17.69
35,000 – 49,999	5.98	23.37	36.61	24.76	9.28
50,000 – 74,999	8.16	28.01	37.11	19.11	7.61
75,000 – 99,999	8.66	33.62	38.71	13.07	5.94
100,000 +	13.12	36.73	32.56	13.70	3.89

Based on the above tables, we noticed that nearly equal number of respondents attributed themselves as having "Excellent" health as did "Very Good," while people were more likely to perceive themselves as having "Fair" health than "Poor." In order to understand if the ambiguity of the cutoffs had an effect, we therefore examined accuracy of the respondents being within ±1 health category (i.e if true health was "Good," they selected "Very Good," "Good," or "Fair") (Table 7). As expected, accuracy increased across all groups as many people may be unaware of the differences between different categories. As seen in Table 4, increasing income still resulted in increasing accuracy for those belonging to "Excellent" health. Furthermore, the opposite effect was again seen for those belonging to "Poor" or "Fair" health. For "Very Good" and "Good" health, differences in income categories were much smaller, suggesting that the effect of income may be much more pronounced at the extremes.

Table 7: Accuracy of Perceived Health Choice within ±1 Category by Income Category and Actual Health Quantile

Income Category ($)	Excellent (%)	Very Good (%)	Good (%)	Fair (%)	Poor (%)
≤ 34,999	59.03	81.79	80.66	63.84	50.93
35,000 – 49,999	66.01	88.80	86.09	48.87	34.04
50,000 – 74,999	69.88	91.66	84.11	42.83	26.72
75,000 – 99,999	73.56	94.09	85.67	36.54	19.02
100,000 +	81.45	96.47	78.18	26.27	17.59

When examining for differences within income groups, we used the same approach as in Table 7 (less restrictive definition of being "correct") to measure accuracy. In Table 8, each row corresponds to the distribution of perceived health accuracy (within ±1 category) across each individual income group, with the row total corresponding to the overall accuracy for each income category. We again see the same trends as noted in the previous two tables of increasing income resulting in increasing accuracy for those with "Excellent" health as well as decreasing income resulting in increasing accuracy for those with "Poor" health. Additionally, when comparing the total accuracy for each income group, we found no major differences between categories except between the lowest and highest income group (~3.5% difference). The only major differences were occurring in which perceived health category respondents were found to be most accurate.

Table 8: Accuracy of Perceived Health Choice within ±1 Category by Income Category

Income Category ($)	Excellent (%)	Very Good (%)	Good (%)	Fair (%)	Poor (%)	Row Total (%)
≤ 34,999	10.96	22.22	6.52	11.29	14.52	65.51
35,000 – 49,999	13.76	27.79	6.25	8.84	7.66	64.30
50,000 – 74,999	15.18	31.57	5.95	7.75	4.98	65.43
75,000 – 99,999	16.05	34.35	6.29	7.15	2.81	66.65
100,000 +	18.95	37.98	4.97	5.02	2.09	69.01

Model Validation

To understand if the observed patterns could be considered legitimate, we performed cross-validation to test the model’s predictive accuracy. To do this, we trained a model using the same parameters described above on 80% of the data and tested its predictive capabilities on the remaining 20% of records. To quantify the predictive accuracy, we first sorted the predictions into quantiles using the same boundaries as the full dataset to resemble the perceived health categories chosen by the participants. Given the frequent inaccuracies described by participants between their health score and their perceived health in ±1 categories (i.e., perceiving oneself as having “Excellent” health but really having “Very Good” health according to the score), we chose to determine the predictive accuracy of our model by assessing the percent of instances where the model predicted a score within 1 category of the true value. With this setup, we ran our cross-validation 100 times, so the figures presented in Table 9 are the averages of these trials.

On average, our model has a 66.67% accuracy rate across all income and health score levels. To assess the usefulness of our engineered score feature and to understand the strength of the relationships among the included variables in our model, we compared it to both naïve and reduced models. A naïve classification system with this data would be one in which every record was predicted to be the health score category that occurs most frequently. This would be the "Very Good" category, which appears approximately 33% of the time. Combining this with the categories adjacent to it, as per our defined prediction accuracy metric, would include 61.83% of records, thus displaying worse performance than our model. Because of the increased performance when using our engineered score feature, this suggests that the new variable is indeed useful in addressing our research questions despite its potential limitations. In a similar manner, we compared our model to a reduced model, which only predicts health score using income category, to better understand the relationship between perceived health and true health and to gauge the extent to which our model outperforms a simpler model. As the prediction accuracy for the reduced model was 60.69%, in comparison to the 66.67% of our model, this suggests both the existence of a relationship between perceived health and true health and an increased performance of our model in comparison to simpler ones.

Given the goals of this exploration, it is important to explore how prediction accuracy changes across different income and health categories. At a general level, Table 9 displays the prediction accuracy for each income category:

Table 9: Model Prediction Accuracy by Income Category

Income Category ($)	Prediction Accuracy (%)
≤ 34,999	64.87
35,000 – 49,999	62.67
50,000 – 74,999	66.75
75,000 – 99,999	67.02
100,000 +	70.29

As evident from this, as income increases, the prediction accuracy also increases. This increase mostly parallels the trend found in the accuracy of participants' own perception of their health versus their true health score ("Row Total" column of Table 8). That these figures are only slightly greater overall than the participants' perception, particularly as income increases, makes sense: previous discussion of the model suggested that the inclusion of income as a predictor increased explained variability by ~2%, so one would expect that the accuracy of our model would be slightly better than participants' own accuracy in assessing their health given this increased information.

Finally, to further understand how our model performs, we stratified the prediction accuracy by both income and score categorization. In Table 10, each cell represents the percent of people in a given category whose health score was accurately predicted (i.e., 29.25% of people making at most $34,999 whose true health score was "Excellent" were accurately labelled based on the model's prediction).

Table 10: Model Prediction Accuracy by Income Category and Health Score Quantile

Income Category ($)	Excellent (%)	Very Good (%)	Good (%)	Fair (%)	Poor (%)
≤ 34,999	29.25	47.63	90.22	89.42	81.99
35,000 – 49,999	30.5	59.23	95.98	83.41	69.62
50,000 – 74,999	37.22	91.71	97.22	83.74	26.97
75,000 – 99,999	40.22	94.17	99.3	70.11	18.71
100,000 +	81.57	96.43	99.97	26.36	17.64

This stratification of prediction accuracy again showcases a similar pattern to that found in the participants' own estimation of their health. Specifically, prediction accuracy for "Excellent" health increases as income increases, and correspondingly prediction accuracy for "Poor" health decreases as income increases. This trend also exists to some degree in the "Very Good" and "Fair" columns, although to a lesser extent than their respective extremities. As discussed in the previous section, this seems to suggest that lower income people are better at identifying their status as being of "Poor" health, and likewise that people of higher income levels are able to better identify their "Excellent" health status; given that this pattern existed in the data, it makes sense that it reappears in the validation. Importantly, prediction accuracy for the "Good" column appears quite high, likely as a result of the way in which we operationalized our accuracy, namely to consider predictions within one level of health score as accurate. That this inflated accuracy does not exist at the extremes suggests a stronger relationship between these levels of health and income.

Comparison of Findings to Existing Literature

In addition to validating our findings by quantitatively assessing our model's accuracy, it is important to also compare these insights to existing literature on the relationship between health perception and income. Key among these patterns is that lower income individuals seem to more accurately perceive their health as being "Poor" than higher income individuals. A possible explanation for this pattern derives from a 2006 interview-based study which found that lower income individuals are more aware than higher income individuals that a relationship exists between their health and their income. People with less income described feeling limited by their finances when trying to make healthy decisions, leading to frustration and the perception that the "divided society" would "take time off [their lives][6]." In the context of our research, this may suggest that, because lower income individuals are more keenly aware of the negative impact of their finances on their health, they tend to be more accurate than higher income individuals in perceiving this reality.

An interesting opportunity for further research could explore why the converse is true, that lower income people are less accurate than higher income people at labeling themselves as having better health. Continuing with the explanation from the previous case, this could be because lower income people may expect that their health would be worse than it actually is due to their low income status. An alternative hypothesis that other research could explore is how such phenomena function for higher income individuals. For example, it could be possible that, since having a higher income probably enables people to both make healthier decisions and have

[6] Davidson, Rosemary, Jenny Kitzinger, and Kate Hunt. "The Wealthy Get Healthy, the Poor Get Poorly? Lay Perceptions of Health Inequalities." *Social Science & Medicine* 62, no. 9 (May 1, 2006): 2171–82. https://doi.org/10.1016/j.socscimed.2005.10.010.

better access to healthcare, people of a higher income with poor health do not perceive it as such. It is possible that they either perceive themselves as healthier given their greater access or view their poorer health less negatively because they have the access to improve it. While our research certainly does not provide evidence to support or refute such hypotheses, it raises questions regarding why these extraneous patterns exist in our data.

Data Snooping and P-Hacking

Throughout the analyses described above, there were some opportunities at which we could have taken advantage of our data to build a model that provides greater predictive accuracy at the expense of interpretation and integrity. For example, when creating our score variable, we chose to include every health condition available to us in composing it. At this point, we could have iterated through these conditions to determine which decreased the predictive accuracy of our model, but doing so would come at the expense of building a model that does not accurately reflect the true health of the survey participants.

Conclusion

Through this investigation, we have uncovered numerous intriguing trends regarding the relationship between perceived and actual health across income categories. Most importantly, our results suggest that, as income increases, individuals become more accurate in identifying themselves as having better health while simultaneously becoming worse at identifying themselves as having poorer health. In part, this seems to be because, regardless of actual health, as income increases, people are more likely to label themselves as having “Very Good” or “Excellent” health, resulting in both many instances when these people are accurate, but also many instances when they are inaccurate when their true health is worse. Likewise, as income decreases, people become increasingly accurate in labelling themselves as having worse health, which corroborates research regarding the impacts of low income on the perception of one's health.

In building and validating a model both to understand the relationship between perceived health, income, and actual health and to verify the usefulness of our engineered health score feature, we determined that our model is more accurate in its predictions than reduced and naïve models, predicting 66.67% of cases accurately. The increased accuracy provided by our model in comparison to other models suggests that our score feature is indeed useful in understanding the relationships that exist among these variables. Despite this seeming usefulness, the score value incorporates some level of uncertainty in our model as a result of potential flaws in its generation, including the averaging of disease severity and the overall lacking breadth regarding the diseases that were included in imputing it. Considering this, the extent to which it truly represents a participant's health adds uncertainty to our model and, as a result, to our findings overall.

In understanding the usefulness and application of these results for public health researchers and strategists, this analysis offers two key findings and some potential avenues for further research. First, public health officials could use the finding that higher income individuals are worse at identifying themselves as having increasingly poor health to spread awareness for members of this population to both more accurately assess their health and take care of themselves. Likewise, this research corroborates the understandings that lower income individuals both perceive themselves as having and actually have worse health than people of a higher income, which is a critical trend discussed when raising awareness about and seeking to change the effects of economic disparity on health outcomes. Finally, while our analysis suggests that these trends exist, they do not necessarily answer *why* such trends exist, namely why higher income people do not accurately perceive themselves as having poor health and vice versa for lower income individuals, thus providing a potential opportunity for researchers in this field to contribute to the existing literature and campaigns that explore and combat the negative effects of income inequality on public health.

Appendix

Peer Critique

Note: The version of Isaac and Begum's report that we received was a work-in-progress with the findings and methodological sections still being fleshed out, so some of the critiques mentioned below may not apply to the final draft of the assignment.

Isaac and Begum's report is most successful both in its use of secondary literature to support its claims and in its interpretation of their findings and analyses. Regarding the former, the students substantiate their methodological background and findings with significant evidence from related literatures, such as their adoption of keywords for depression and anxiety from previous studies of social media. Doing this provides strong background for the students' approach in learning about the prevalence of such expressions of mental health disorders among Twitter users. Likewise, when interpreting their findings, they used interesting social observations to develop hypotheses for unexpected insights, thus providing interesting avenues for further research. This was exemplified well in their findings related to gender, as numerous sociopsychological factors may influence the different ways in which males and females present themselves both online and in person, so noting potential explanations for differences in their findings was important both in the framing of their analyses and in enabling their research to be used for further investigations.

Alongside these successes, the group has some opportunities in their paper where their focus and methodology could be explained in further detail to increase the clarity of the project's aims and approaches for the reader. For example, in their introduction, while the goal of learning about the expression of mental health on Twitter is evident, the exact guiding question is not made totally clear. While the goal seems to be to "understand the prevalence and nature of depressive and anxious thoughts on Twitter," the conclusion of the report does not provide a direct response to this topic, leaving the guiding question somewhat ambiguous by the end of the report. Moreover, some of the mathematical and methodological steps taken by the students could be discussed in a more detailed fashion to ensure that the steps taken throughout this research are reproducible. For example, while the students discuss their use of a TF-IDF data structure, they do not detail how they tokenized each tweet or what TF-IDF method they used in determining each token's weight. The group did have some success in detailing their transformation and engineering of certain features, though, such as in the well-articulated description of the tweet ratios for males and females. Continuing this level of detail throughout the report would enable readers to better understand the decisions made both in structuring data and building models to address their research questions.

Considering these elements, Isaac and Begum display a strong understanding of the literature in which their research exists and how they could use it to generate further research questions. However, alongside the implementation of such literatures, the group has room to improve in their detailing of the project’s focus and the key methodological decisions made in building algorithms and forming data visualizations to explore this topic.

Appendix Table 1: Medical Conditions Used to Engineer Score Feature

Medical Conditions	Asthma, Arthritis, Coronary Heart Disease, Angina, Myocardial Infarction, Stroke, Chronic Obstructive Pulmonary Disease, Abdominal Pain in past 3 months, Cancer types: Bladder, Blood, Bone, Brain, Cervical, Colon, Esophageal, Gallbladder, Larynx-Trachea, Leukemia, Liver, Lung, Lymphoma, Melanoma, Mouth, Ovarian, Pancreatic, Prostate, Rectal, Skin Non-Melanoma, Skin Unknown Kind, Stomach, Thyroid, Uterine